
Analysis of Transcriptional Regulatory Regions

James W Fickett
SmithKline Beecham

Contents

The state of recognizing

- » eukaryotic promoters
- » transcription factor binding sites
- » regulatory regions
active in a particular context

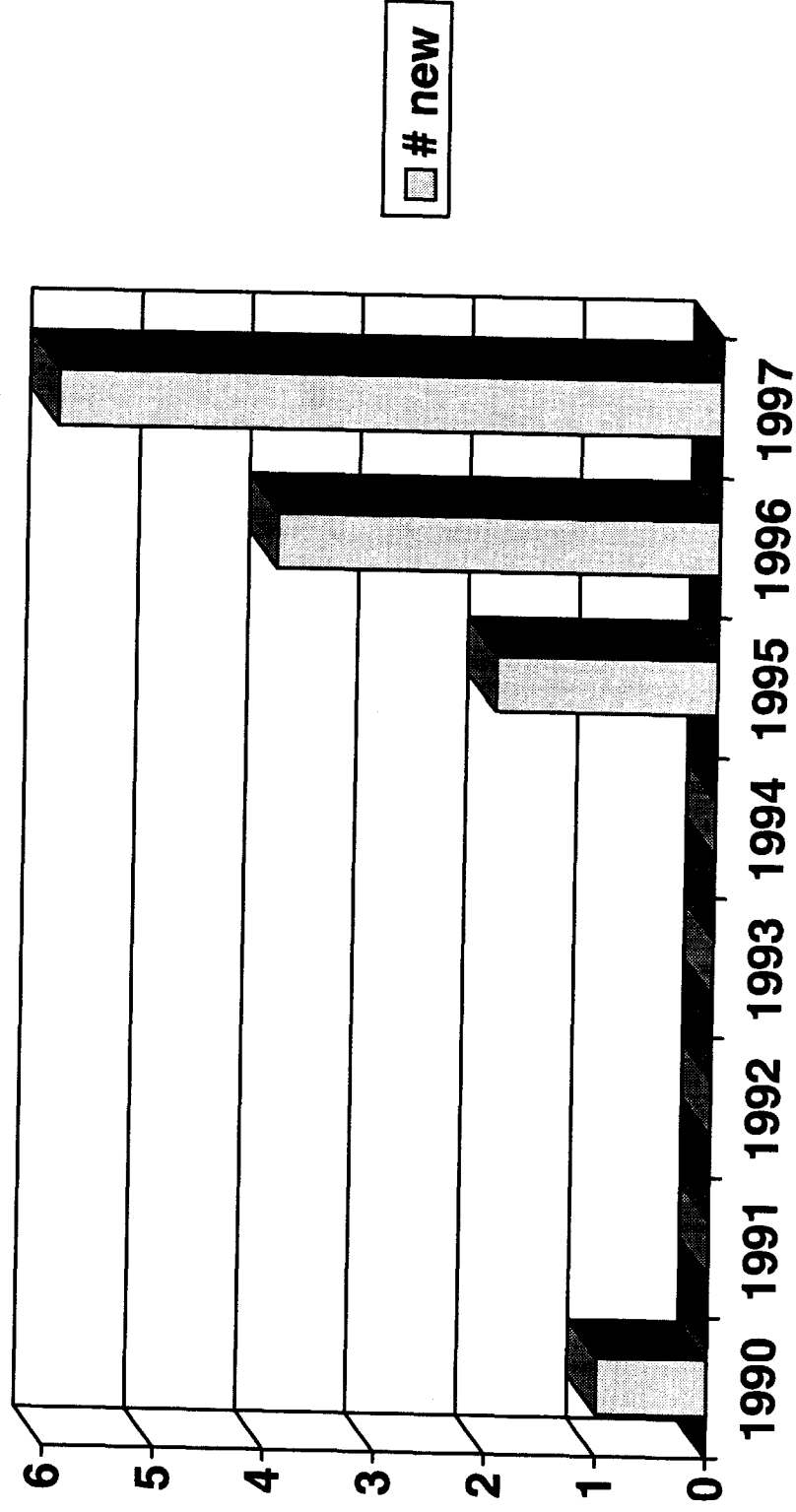
Why promoter recognition?

- Identifying multiple/partial genes
GENSCAN, BCM GeneFinder, (GeneID)
Promoter is one syntactical separator
- One place to look for functional clues

Overview, current algorithms

- What's out there
- How well do they work
- Possible directions

New algorithms



Via clusters of regulatory sites

- Autogene, Kondrakhin &, CABIOS, 1995
- PromoterScan, Prestridge, JMB, 1995
- PromFD, Chen &, CABIOS, 1997
- (Bayesian), Crowley &, JMB, 1997
- TSSG/TSSW, Solovyev &, ISMB, 1997

Improvements possible?

- Large amount of experimental data not being used in current algorithms (guide to biology relevant to computation in Genome Res 7, 861)
- Core promoter elements complex; May need multiple/nonlinear models

Initiator region

- The usual descriptions of initiator specificity contains only about 5 bits of information
- But in a TATA-less promoter the initiator seems to be able to localize transcription initiation within a region of ~hundred base pairs

Via general patterns

- PromFind, Hutchinson, CABIOS, 1996
- Promoter1.0, Knudsen, unpublished
- (HMM), Pedersen & ISMB, 1996
- (Markov), Audic & Comput Chem, 1997

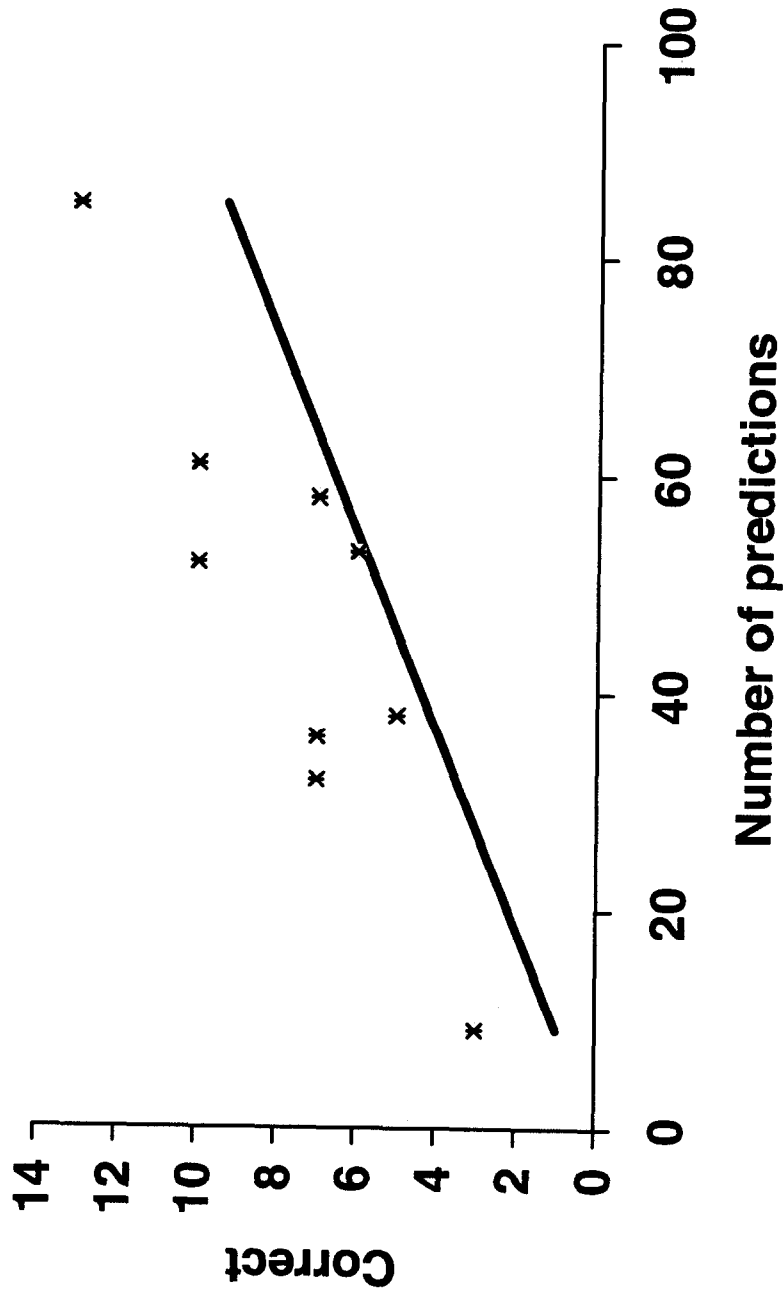
Via core promoter elements

- TATA, Bucher, JMB, 1990
- GRAIL, Matis, Comput. Chem., 1996
- (TDNN), Hatzigeorgiou & IEEE IJISIS, 1996
- NNPP, Reese, unpublished

A small-scale test ...

- ... to get some idea if problem solved
- 24 recently mapped mammalian transcription initiation sites in 18 sequences totaling 33 kb
- “Correct” if -200 to +100

Performance



State of the art

- Current programs correctly predicted 13% - 54% of the promoters...
- ...with on the order of one false positive per kilobasepair
- Many caveats, but still it seems the problem is not solved

Recognizing Protein-DNA Binding Sites

At the heart of most transcriptional analysis

Can we do it?

**In most cases, with a lot of work per protein,
yes**

Protein-DNA Binding Sites

Caveat: Binding site != Active site

E.g. Tronche et al., JMB, 1997. Experimental verified prediction of HNF1 binding. But many predicted sites in genes not regulated by HNF1.

**Probably inactive binding sites are hidden.
Leave this problem aside for now.**

PWMs; scoring a potential site

a	-1.67	-2.86	-2.86	1.84	0.63	1.48	1.65	1.77	-2.86	1.84	-0.97	0.94
c	-2.86	1.65	-2.86	-2.86	-2.86	-2.86	-2.86	-2.86	-2.86	-2.86	-2.03	0.06
g	1.18	-2.86	-2.14	-2.86	-2.86	-2.14	-2.86	-2.14	-2.86	-2.86	1.62	-1.25
t	0.36	-0.79	1.81	-2.86	1.12	-0.24	-0.79	-2.14	1.84	-2.86	-2.68	-0.74
	C	A	A	G	A	T	G	C	T	A	A	A

-2.86 -2.86 -2.86 -2.86 +0.63 -0.24 -2.86 -2.86 +1.84 -0.97 +0.94

= -13.12

- Assume independent positions within site
- Can be thought of as log likelihood ratio (site model vs background model) or predicted energy of binding
- Work pretty well, and *much* better than consensus sequences

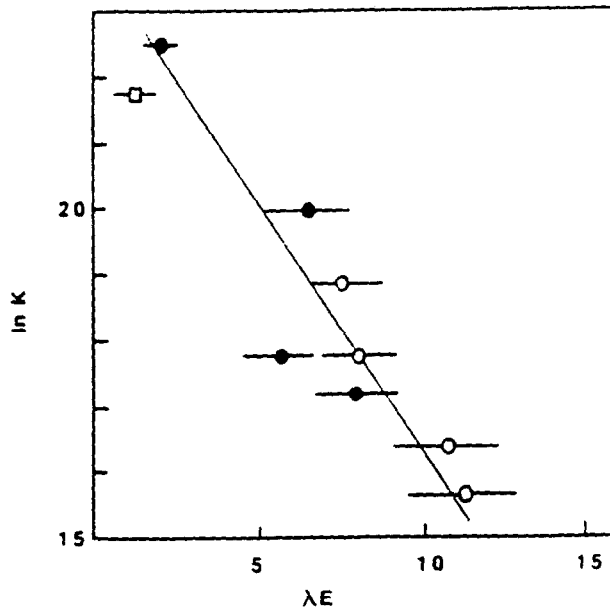


Figure 2. Correlation between the sequence heterology index λE and the logarithm of the binding constant for the LexA protein to single sites from the data listed in Table I. ● refer to *in vitro* binding constants and □ refer to binding constants estimated from the *in vivo* induction ratio relative to the binding at the *recA* site. The four data points indicated by ○ refer to single base-pair mutations of the *recA* operator with induction ratios given by Wertman & Mount (15). The error bars are determined from the small-sample uncertainty in the heterology index (1,2). The correlation line is a least-squares line minimizing deviations in λE .

mately $2 \cdot 10^{-8}$ M; this corresponds to about 12 free *lexA* dimers per cell which is a reasonable value. Actually, the binding constant is not necessarily the same under physiological conditions and this estimate of the free concentrations could be substantially off.

Furthermore, we can consider the induction ratio at the *lexA* operon containing two operator sites. Using $K_0 R_f = 10.5 \exp(6.7/\lambda)$ from the *recA* data discussed above and the heterology indices $\lambda E_1 = 10.3$, $\lambda E_2 = 7.1$ (see Table 1) one finds from Eq. (5) the expected induction ratio:

$$1/P_0 = 1 + 10.5[\exp(-3.6/\lambda) + \exp(-0.4/\lambda) + w10.5\exp(-4/\lambda)] \quad (9)$$

If $\lambda = 1.3$ and $w=1$, this gives $1/P_0=14$ not too different from the observed induction ratio 8.8 (15), thus confirming the expectation that there is no significant cooperativity ($w \approx 1$) between the sites. For the two operator mutations pKW13 (changing λE_1 from 10.3 to 9) and pKW14 (changing λE_2 from 7.1 to 5.8) one finds in the same way $1/P_0=24$ and $1/P_0=36$, respectively.

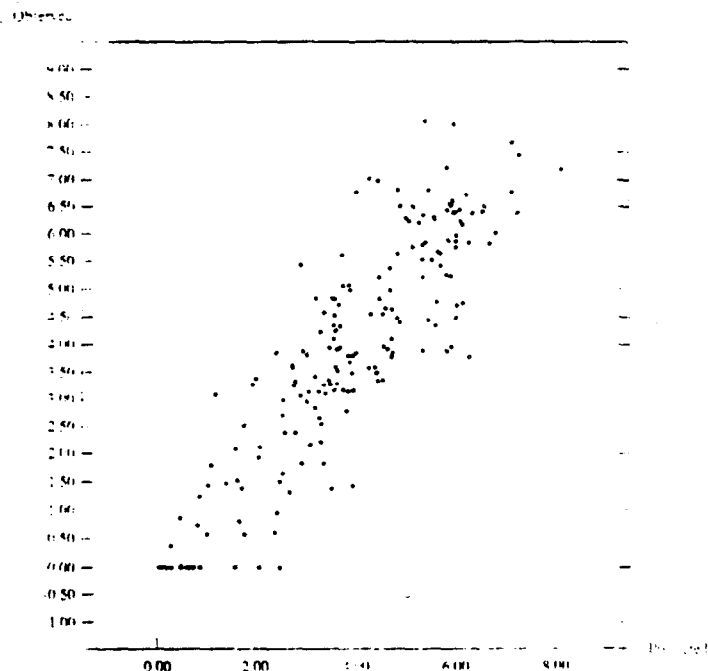


Figure 2. Plot of the predicted versus the observed $\ln(\beta\text{-galactosidase activity})$ values for the 185 ribosome binding sites. The predicted values were determined by matrix evaluation of the corresponding sequences using the matrix in Figure 1b, as described.

of the activity, only 2% of the total variance in activity measurements is due to the variability of the measurements themselves, and the remaining 98% can be attributed to differences in the sequences. This also represents the upper limit of r^2 that we could get from the analysis based on the sequences of the isolates.

Mononucleotide analysis on all data

A regression analysis was performed [essentially as in Stormo *et al.*, (23)] on the entire data set of 1012 specific activities for the 185 sequences. Figure 1 shows the values of the base-position parameters obtained, both in normal matrix notation (23,28) and graphically. Figure 2 is a graph of the observed vs. expected values for each of the 185 sequences, using the matrix of Figure 1b for calculating the expected values. The correlation coefficient between the observed and predicted values is 0.89 ($r^2=0.79$; all r 's are corrected for degrees of freedom). Many of the features displayed in Figure 1 are consistent with the known qualitative aspects of ribosome binding sites. For example, a G-rich sequence in the -6 to -11 region leading to high level translation is consistent with that being a Shine/Dalgarno sequence. The high activity of A at -3 is consistent with the statistics for that position (2), and some activity measurements (13). The fact that A at position 0 leads to the highest expression is also as expected, as it is that C at position 0 gives the lowest expression (or even works at all). However, some other features are surprising. For example, we expected G at 0 to be better than T (12). We also did not expect a T at -11 to lead to nearly as high expression as a G, and better than an A. These two features, in particular, can be explained by postulating that translation may begin at positions other than 0. In order for functional $\beta\text{-galactosidase}$ to be produced, translation must begin in the same frame as 0, such as -3 , -6 , -9 or -12 . In fact, position -12 is always

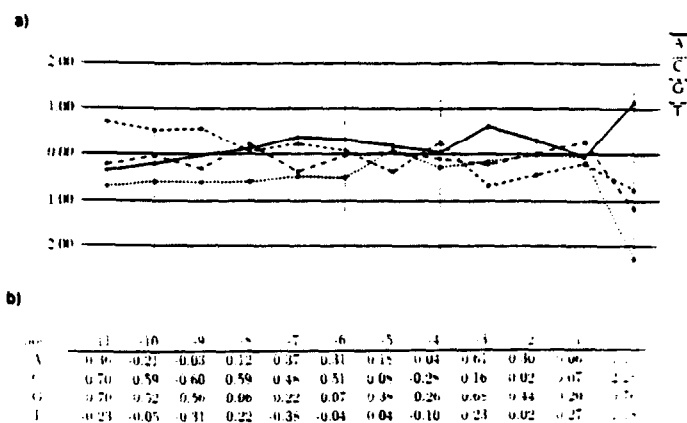


Figure 3. a) Normalized values of base versus position for mononucleotide analysis of 158 ribosome binding sites without alternative ATGs. b) The same values in normal matrix form, as in Figure 1. The constant in this case is 4.64.

an A, so whenever -11 is a T and -10 is a G (10 such cases, out of 42 Ts at -11), translation may be higher than expected due to in-frame initiation at this position. This could account for T at -11 leading to high levels of translation. Similarly, whenever there is a G at position 0 there is some probability that it is preceded by an AT (7 such cases, out of 51 Gs at position 0), which could lead to initiation in an alternative reading frame. Those cases would decrease the average contribution of a G at position 0 to the activity. Since the amino-termini of the $\beta\text{-galactosidase}$ proteins were not sequenced we do not know the initiation codon that is actually used.

Mononucleotide analysis excluding alternative ATGs

Several analyses were performed to assess the effect of alternative initiation codons, and to determine the mononucleotide parameters in the absence of those effects. In the first analysis all of the sequences with ATG occurring in positions other than 0 were simply eliminated from the data set. This removed 27 sequences and a total of 140 activity measurements. The analysis described above was repeated on this smaller set of 872 activity measurements on 158 sequences. (In a separate analysis we also removed alternative GTG sequences, since they should have the second largest effect, but no additional improvement was seen; data not shown). Figure 3 shows the parameters obtained, and the filled squares in Figure 4 plot the observed vs. expected activities for this matrix and this set of data. The correlation coefficient is 0.92 ($r^2=0.85$), a substantial improvement over the previous analysis that demonstrates the effect of alternative ATGs upstream of the initiation codon. Wild-type initiation sites rarely have alternative ATGs in the vicinity of the initiation codon (1,2). The standard deviation of the difference between the predicted and observed values is only 0.82, indicating that most of the time the observed activities are within about 2-fold of those predicted by the matrix in Figure 3b.

Removing the alternative ATGs from the data did not significantly change most of the matrix values, but did make G much better than T at position -11 . This analysis also separated G and T at position 0, with G being better, as expected. In fact, the values obtained for the relative activities for ATG, GTG and TTG initiation codons, 1, 0.15 and 0.10, respectively, are identical to the relative rate constants for the different initiation

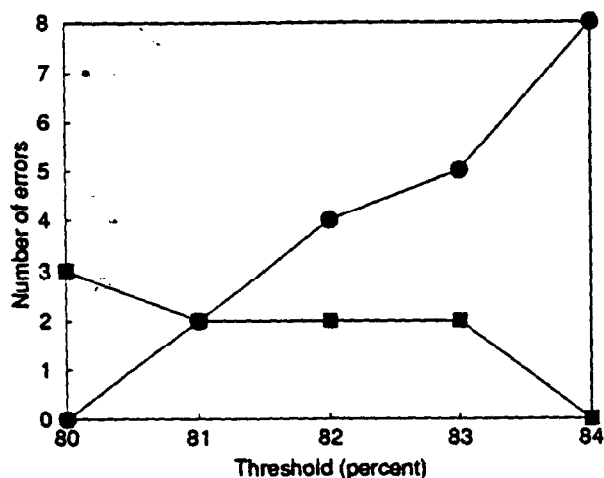


FIG. 3. Accuracy of MEF2 site recognition with the PWM score. Symbols: ○, number (out of 22) of natural sites missed at each threshold; □, number of apparent false-positive sites in the 400-bp neighborhoods of the natural sites. Predicted sites overlapping true sites by at least 10 of 12 bases, on either strand, were not counted as false positives.

threshold on both strands, it was counted only once); 21,199 sites were found. This gives an upper bound of one false-positive site per 1.7 kb. Since some of the sites so found are, of course, true functional sites, the false-positive rate is probably more like one site per 2 kb for this set of data.

Other discrimination methods seem to work less well. It is not obvious a priori whether a PWM built from binding site selection data would be better or worse, for recognition of natural MEF2 sites, than one built from the natural sites themselves. The binding site selection data presumably represent purer information, in that only the binding affinity of a single protein is involved, while in the natural sites there are probably functional constraints other than binding affinity and the possible involvement of several forms of MEF2. The idea of using a natural-site PWM was tested as follows. Since it is desirable to use the PWM to recognize new sites, circularity was avoided by choosing each of the 11 sites in turn, building a PWM from the other 10, and then testing it on the one left out (the so-called "jackknife" technique). The results are shown in Fig. 4. Although the 83% threshold is somewhat attractive, it appears that the PWM from the binding site selection data gives cleaner discrimination.

The use of consensus sequences sometimes represents less a rigid mechanical algorithm than a mnemonic for the application of considerable specialized knowledge. Nevertheless, to gain some idea of whether use of the (admittedly more complex) PWM gives a return of some value, one implementation of the consensus approach was evaluated as follows. In the frequency matrix from the binding site selection data, any base gaining more than 10% of the count was included in a consensus, to give the pattern (g/t)(c/t)ta(a/t)(a/t)ata(a/g)(a/c/t). A search requiring a perfect match resulted in no false positives but missed 12 of the 22 sites. When a search was done with one mismatch allowed, only two sites were missed but there were nine false positives. A search with two mismatches allowed picked up all of the true sites but resulted in 21 false positives.

PWM scores order sites roughly according to strength. When the 400-bp neighborhoods of the natural sites were scored with the PWM from the binding site selection data, the natural site in each sequence was always the highest-scoring

site, even when other, apparently false-positive, sites were located. To investigate further whether the PWM scores rank sites in rough accord with activity, data on relative binding strength, or relative effect on expression, of multiple wild-type or mutated natural sites were collected and the ranking of the sites by PWM score was compared with the ranking by experimental measurement (Table 1). The relative rankings of sites across sets of data derived from different experiments cannot be compared, but within each set of data the ranking based on PWM scores is in rough accord with the ranking based on experimental data. When a lower-ranking oligomer has a higher PWM score than a higher-ranking one, the scores are always close.

Oligomer mt6 is of particular interest, as it has been shown to differentiate a ubiquitous factor with similar specificity, which binds mt6, from MEF2, which does not (see references 10 and 47 for MEF2A, 25 and 31 for MEF2C, and 5 and 30 for MEF2D). However, the mt6 score, which is 82%, is borderline and at the same level as some other false-positive sites (Fig. 3). (The PWM from the natural sites also incorrectly ranks some of the oligomer data and ranks the mt6 site above some sites with positive MEF2 binding [data not shown]).

The PWM may help in sorting sites of factors with DNA-binding specificity overlapping that of MEF2. The A/T-2 site in the α -cardiac MHC gene has a score of 80%, a score inferior to that of any known MEF2 site. Thus, had the PWM been available when this gene was studied, it might have speeded the discovery that this site is activated by ARF rather than MEF2 (34). Of course, not all cases are so easy. The MLC2 site, which was suspected to interact with MEF2 but is, in fact, an HF1b site (48), has a score of 88%. Although some MEF2 may well bind the HF1b site in vivo, "antibodies directed against the HF-1b zinc finger domains can remove the major component of endogenous HF-1b binding activity in cardiac extracts, while these antibodies have little effect on the MEF-2 binding activity in skeletal muscle cell extracts" (48). The specificity of binding to this site may well have more to do with cooperation between HF1b and the ubiquitous factor that binds the adjacent HF1a site than with the HF1b site itself. A similar comment applies to the A/T site in the MCK enhancer, where it is known that MHOx binds more strongly than MEF2 but it is MEF2 that activates the gene (9).

A recent report by Grayson et al. (17) showed that the A/T

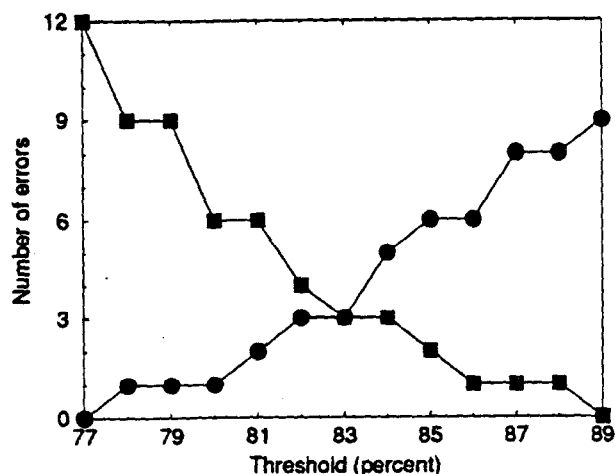


FIG. 4. Accuracy of MEF2 site recognition with the natural-site PWM score. Symbols: ○, number (out of 22) of natural sites missed at each threshold; □, number of apparent false-positive sites in the neighborhoods of the natural sites.

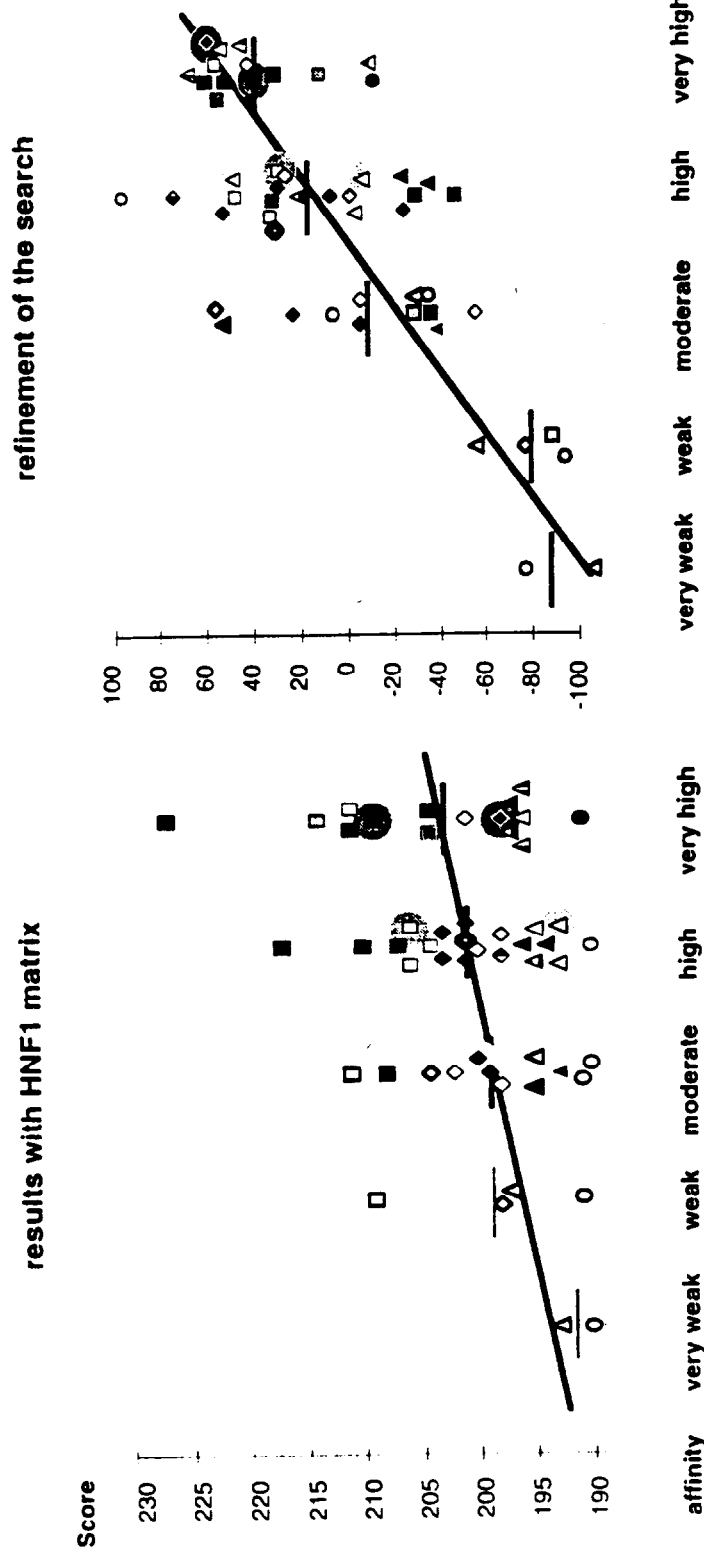
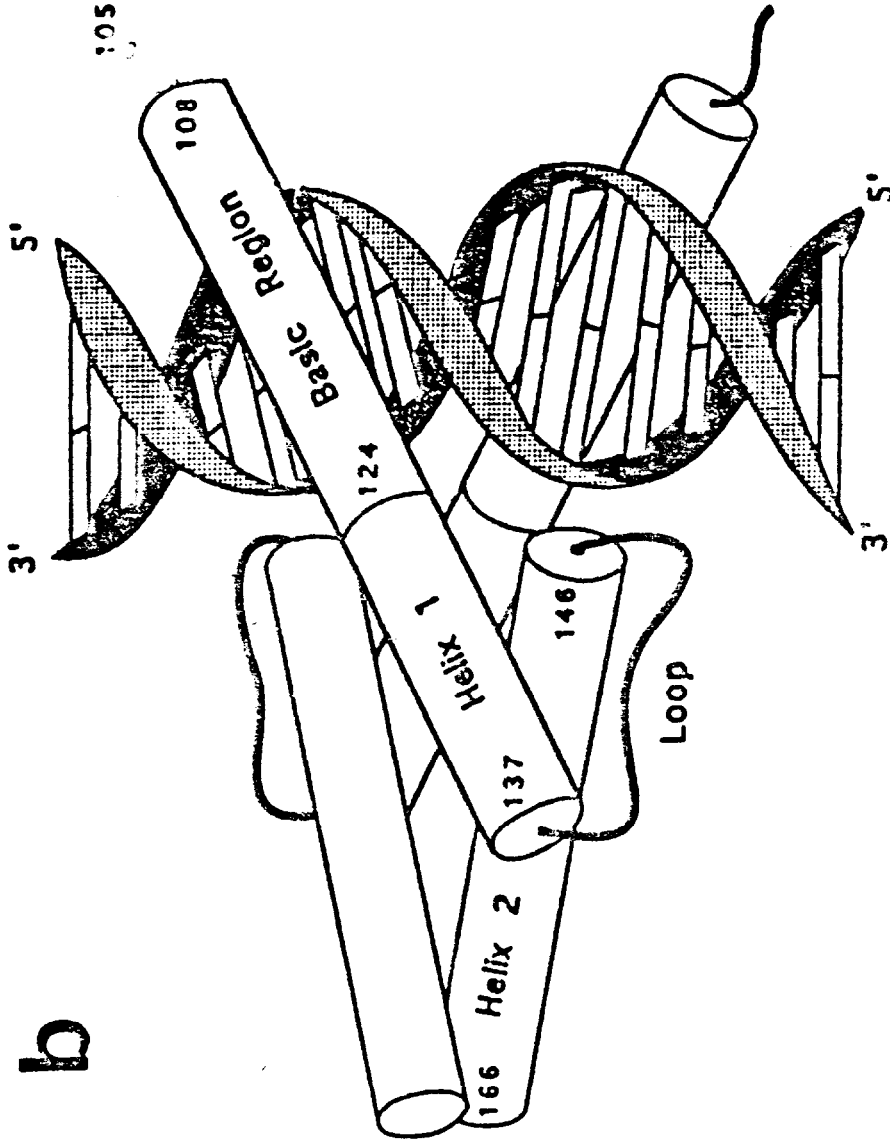


Figure 4. Correlation between the score and the affinity for HNF1. Plot of the affinity of HNF1 protein sequence (abscissa) against the score obtained with the weighted matrix (ordinate). The distribution obtained the HNF1 matrix is given in the left panel. As detailed in the text and in Figure 3, we observed a bias in the nucleotide composition of the sequences depending on their affinity for HNF1. To take this bias into account a new matrix obtained by subtracting the value of a matrix generated with weak and very weak affinity HNF1 sites from the value of the one generated with very high affinity sites. The affinity/score relation obtained this second matrix is shown in the right panel. The grey ellipses indicate the sequences for which we performed a Scatchard analysis. Compared to the affinity of HNF1 for the rat albumin promoter site (PE 56 high affinity sequences (vitamin D binding protein and sucrose isomaltase promoters, classified as very high affinity) have a 15-fold higher K_d . The site present in the corticosteroid binding globulin promoter ("high affinity") has a K_d 15-fold lower. The sites present in the P450IIE1 gene and the tissue plasminogen activator genes (both with a "moderate affinity") have K_d s 1.6 and 11-fold lower, respectively.

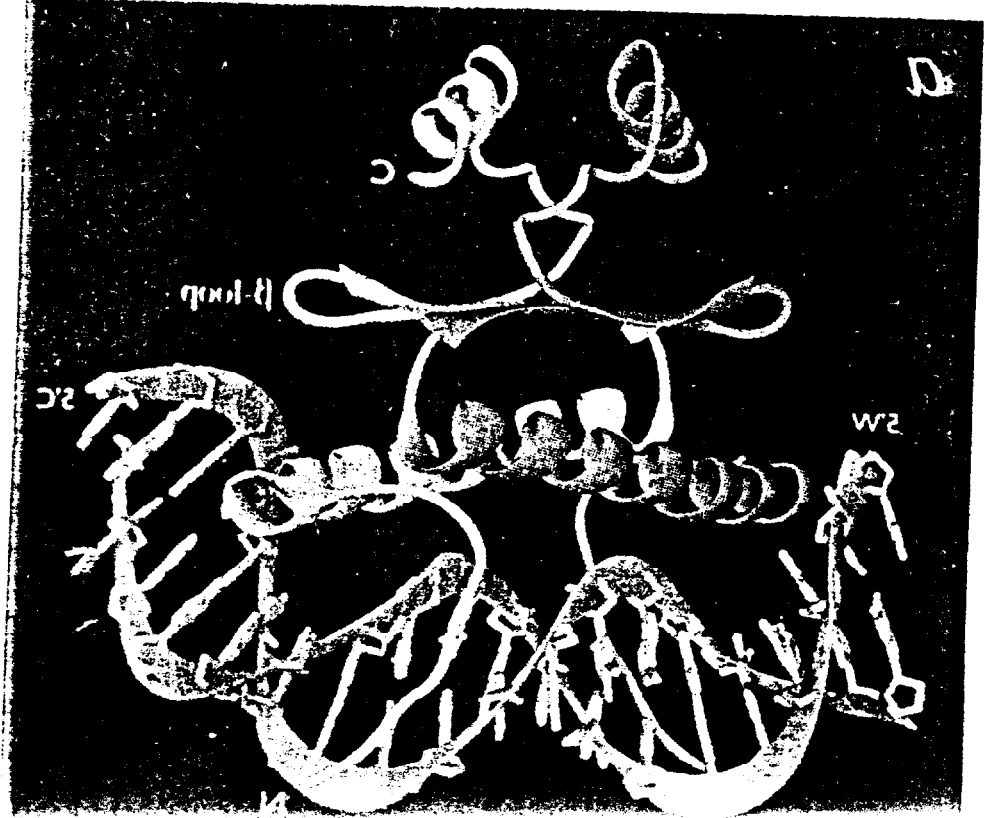
the sites tested for binding (Figure 4). This improvement of the correlation between score and relative affinity shows that at least part of the

HNF1 binding sites are present throughout the genome, but are enriched within live specific genes

b



(which appears at the corresponding position in Max) knocks out DNA binding (Van Antwerp et al., 1992). Max is rather different in this region: there are no contacts with the flanking bases, binding site selections show no sequence preferences at these positions (Blackwell et al., 1993), and the leucine at position 117 does not make a



E-Box Variants Direct Formation of Distinct Complexes with the Basic Helix-Loop-Helix Protein ALF1

Bjarne J. Bonven*, Anders L. Nielsen, Peder L. Norby
Finn S. Pedersen and Poul Jorgensen

Department of Molecular
Biology, University of
Aarhus, C. F. Møllers
Allé 130, DK-8000
Aarhus C, Denmark

The murine transcription factor ALF1 belongs to the class of basic helix-loop-helix proteins specific for the NCAGNTGN-version of the E-box. Binding of homodimeric ALF1 to variants of this motif was studied by a combination of binding site selection technology and DNA modification interference analysis. The results showed that substitutions at the non-conserved positions in the E-box sequence could cause profound alterations in the patterns of specific contacts at the protein-DNA interface. Thus, both the overall extent of the binding region and the backbone phosphate contact pattern differed markedly between closely related E-boxes with similar affinities for ALF1. The identity of the base at the inner N was an important determinant of contact pattern specification. The E-box variants differed in their ability to mediate ALF1 dependent transcriptional activation *in vivo*. We discuss the possibility that adaptability in basic helix-loop-helix protein-DNA interactions can result in complexes with different functional properties.

*Corresponding author

Keywords: ALF1; bHLH protein; E-box; backbone contacts

Introduction

The basic helix-loop-helix (bHLH) transcription factors are involved in transcriptional control of a variety of tissue-specific and developmentally regulated genes (Murre *et al.*, 1989a,b; Zhuang *et al.*, 1994; Bain *et al.*, 1994; Sun, 1994; Guillemot *et al.*, 1994; for a recent review, see Murre *et al.*, 1994). A general picture of the structure of these proteins and their complexes with DNA has emerged from recent work (Ferre-D'Amaré *et al.*, 1993, 1994; Ellenberger *et al.*, 1994; Ma *et al.*, 1994). The helix-loop-helix (HLH) part of the bHLH structural motif is composed of two amphipathic helices connected by a loop. A conserved basic region located immediately N-terminal to the HLH region mediates sequence-specific DNA-binding. The HLH region is a dimerization interface serving to orient the basic regions of the monomers correctly with respect to DNA-binding (Murre *et al.*, 1989a; Voronova & Baltimore, 1990). When docked into DNA, the basic regions lie in the major groove on either side of the DNA straddled by the dimer. Amino acid residues in the basic region are engaged in specific contacts to bases in the major groove and to backbone phosphate groups. The otherwise unstructured basic

region adopts α -helicity upon DNA binding, such that it forms a continuous α -helix with the neighboring helix 1 of the HLH region (Starovasnik *et al.*, 1992; Ferre-D'Amaré *et al.*, 1994). The two amphipathic helices within each monomer cross one another at the loop region and form a four-helix bundle in the dimer. Helix 2 from each monomer further stabilizes dimerization by coiled-coil interactions. Thus, the dimer can be pictured as a globular four-helix bundle from which two pairs of α -helices extend in opposite directions (the basic region/helix 1-pair extend into the DNA and the helix 2-pair extend away from the DNA). The dyad axis of the dimer forms a 90° angle with the helical axis of the DNA. The N-terminal base of helix 2 is closely apposed to the DNA and contacts specific backbone phosphate groups (Ferre-D'Amaré, 1993, 1994; Ellenberger *et al.*, 1994; Ma *et al.*, 1994). A leucine zipper extending helix 2 to form the bHLH/ α -motif is present in a wide range of eukaryotic transcription factors including the Myc-subfamily (Murre *et al.*, 1989a; Blackwood & Eisenman, 1991).

Most bHLH and bHLH/ α proteins recognize a hexameric binding site, CANNTG, known as the E-box. Each monomer contacts bases in only one half-site. The binding specificity is mediated by a conserved glutamic acid/arginine pair in the basic region (Dang *et al.*, 1992; Van Antwerp *et al.*, 1992; Fisher & Goding, 1992; Prochownik & Van Antwerp

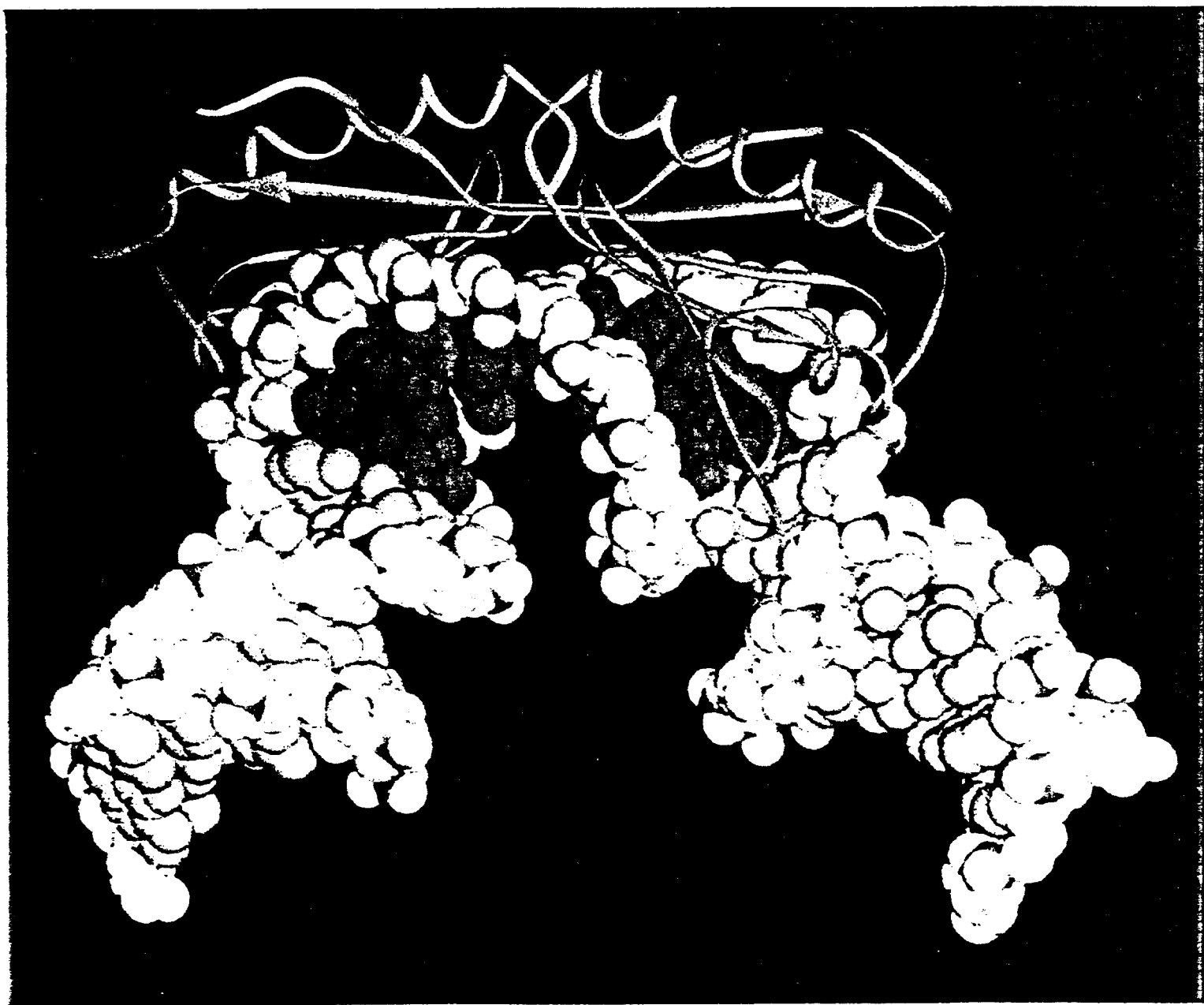
Non-linear Description of TBP Binding Needed?

IIB MyHC CTATAAAG -> tTAatAtAG

Mutant still binds TBP strongly, but with different DNA contacts

Enhancer function is lost

(Diagana et al. JMB 1997)



The big problem is data

There are probably thousands of transcription factors yet to be discovered

Even when the factor is known, often only one or two sites are known

Even when many binding sites are known, careful reading of many papers is required to separate out the sites of proven binding and proven function

Function of new genes

- Systematic ORF knockouts in yeast
- High-throughput mRNA vs tissue/state
- Computer analysis of putative proteins

- Analysis of transcriptional regulatory regions?


Regulatory region analysis

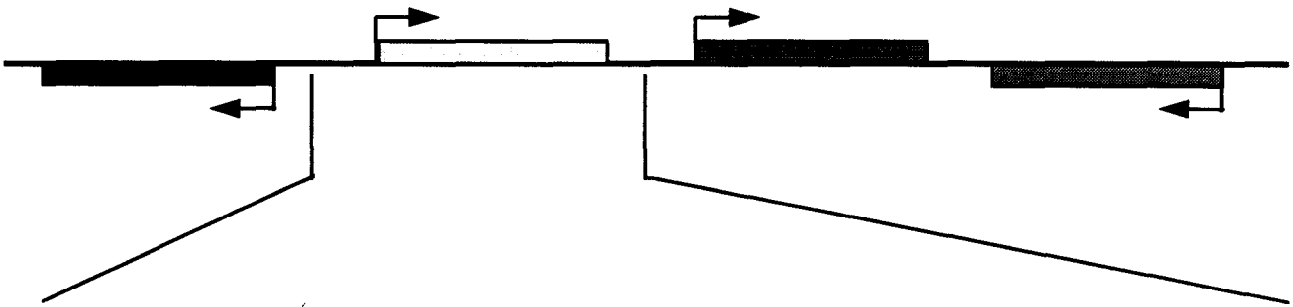
- Claverie & Sauvaget 1985
Significance of spaced consensi
- Fondrat & Kalogeropoulos 1994
Yeast prom. governed by one factor
- Wagner 1997
Clusters of consensus occurrence
- Frech, Quandt & Werner 1997
Grammars of LTR classes

A Starting Point

Develop a computer program that can analyze 200 bp of DNA sequence, and say whether or not it contains a transcriptional regulatory region able to drive skeletal muscle-specific expression

Computational Identification of Regulatory Sites

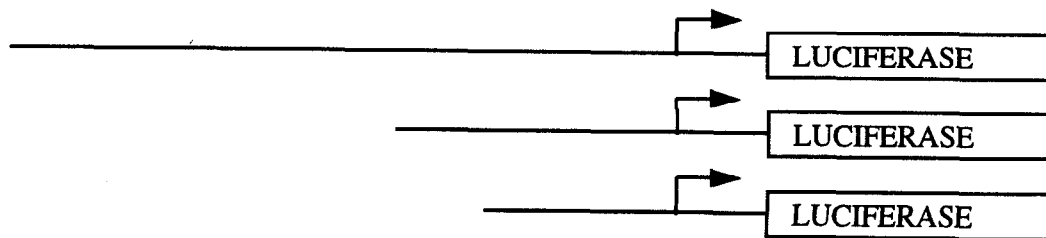
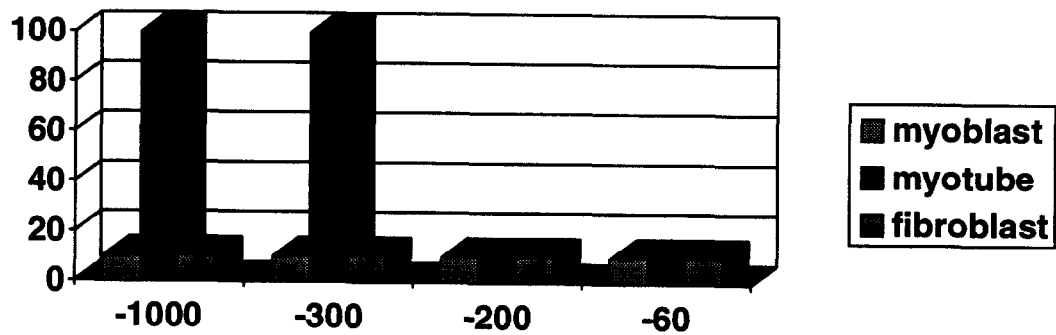
- ◆ A computational method identifies 60% of muscle regulatory regions
 - ◆ Method is specific
 - ◆ Phylogenetic footprinting complements analysis
- 




Environmental, Physiological, and Developmental Controls




Traditional Promoter Analysis



Desiderata for Computational Regulatory Sequence Identification

- ◆ Based on Biology
 - ◆ Quantitative
 - ◆ Easily Interpreted
 - ◆ Adaptable
 - ◆ Accurate
 - Specificity is the challenge
- 

Profiles/Weight Matrices

- ◆ Quantitative score for potential sites
 - ◆ Align known sites using Gibbs sampling
 - ◆ Determine frequency of each nucleotide at each position
- 



A	10	0	0	0	22	0	6	2	3	4	22	10
C	0	2	12	0	0	0	0	0	0	0	0	0
G	9	20	2	0	0	0	0	0	0	0	0	10
T	3	0	8	22	0	22	16	20	19	18	0	2
Pos	1	2	3	4	5	6	7	8	9	10	11	12



A	7	9	4	0	16	7	0	6	0	0	6	0
C	8	0	2	15	0	0	15	0	0	10	0	0
G	1	7	10	1	0	9	1	0	16	6	0	16
T	0	0	0	0	0	0	0	10	0	0	10	0
Pos	1	2	3	4	5	6	7	8	9	10	11	12



A	0	3	2	1	0	0	4	0	0	0	0
C	6	4	8	10	12	12	0	12	12	12	11
G	5	2	2	1	0	0	6	0	0	0	1
T	1	3	0	0	0	0	2	0	0	0	0
Pos	1	2	3	4	5	6	7	8	9	10	11



A	7	9	0	0	18	9	15	8	21	14	0	0	7
C	5	2	21	17	0	0	1	0	0	2	0	0	8
G	7	4	0	0	0	0	4	2	0	0	21	21	6
T	2	6	0	4	3	12	1	11	0	5	0	0	0
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13



A	1	9	0	12	0	0	0	0	5	1	2	0
C	6	0	12	0	0	0	12	11	0	7	4	2
G	1	3	0	0	0	0	0	0	0	4	3	8
T	4	0	0	0	12	12	0	1	7	0	3	2
Pos	1	2	3	4	5	6	7	8	9	10	11	12



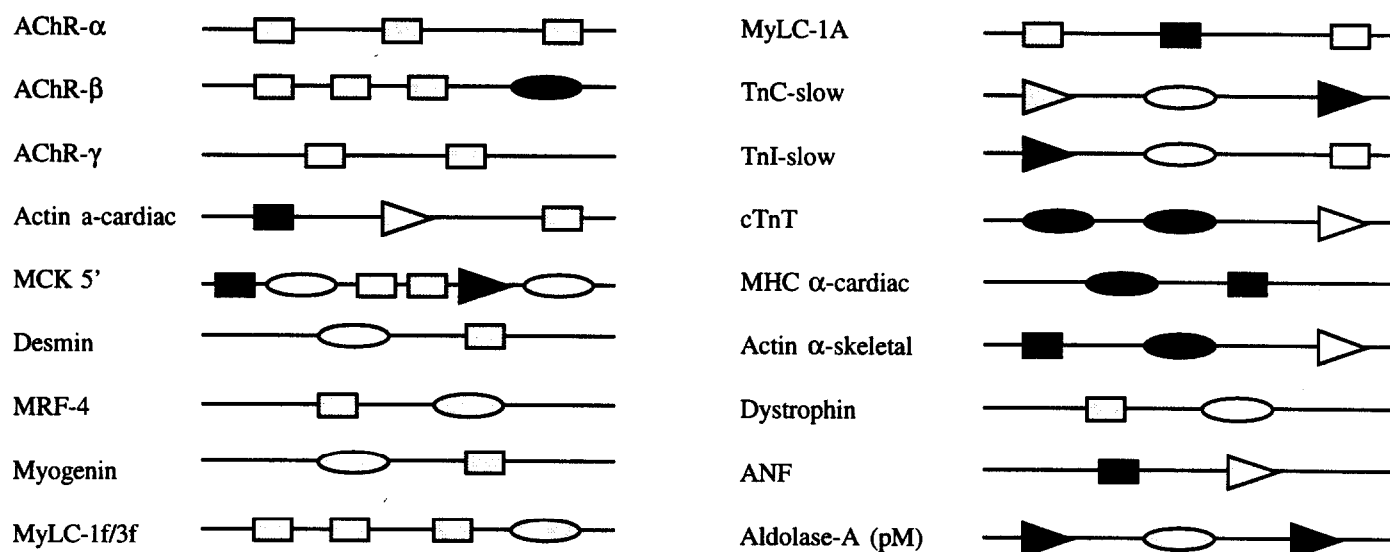
Characteristics of Matrices

<u>Factor</u>	<u>Site Name</u>	<u>#Sites</u>	<u>Scores (%)</u>	<u>BP/Hit</u>
Mef-2	A/T-rich	22	88-100	3700
Myf	E-box	16	82-99	390
SRF	CArG	17	86-100	3000
TEF	M-CAT	21	89-97	7100
Sp1	G/C-box	12	85-99	570

Identification of Muscle Sites

- ◆ 60% of promoter sequences in EPD contain a putative site
- ◆ Inclusion of Sp1 profile raises this to nearly 100%
- ◆ Presence of a site is not predictive for muscle regulatory regions





□ Myf/E-box

■ SRF/CArG box


○ Mef-2

● TEF


△ Sp-1

▲ Novel

Logistic Regression Analysis

- ◆ Commonly used in medical prediction studies
 - Will patient benefit from surgery?
 - ◆ Values output range from 0 to 1
 - Quantitative and Easily Interpreted
 - ◆ Easy to assess contribution from each data field
 - Interpretable
 - ◆ Easy to modify with additional information
 - Adaptable
- 

LRA Training Set

- ◆ 1500 Randomly selected 200 bp sequences from Gb:primate (<1%)
 - ◆ 300 Randomly selected 200 bp sequences from EPD (7%)
 - ◆ 4 sequences containing di- and tri-nucleotide repeats
 - ◆ 29 muscle regulatory sequences of 200 bp
- 

LRA Performance

- ◆ Muscle Regulatory Regions Identified
 - 60% of Test Set Found
- ◆ Predictions are Specific
 - Only 4% of EPD Sequences are Positive



Test LRA with Genomic Sequences

- ◆ Obtained GenBank sequences >200,000bp
 - 11 Human Genomic Regions
 - Masked Repeats
- ◆ 1 Region Found Every 34,000 bp
- ◆ Scrutinized Well Annotated Sequences
 - 21% of Regions Near Muscle Genes
 - Muscle Genes: EMD, DNaseI-like, and HXC26

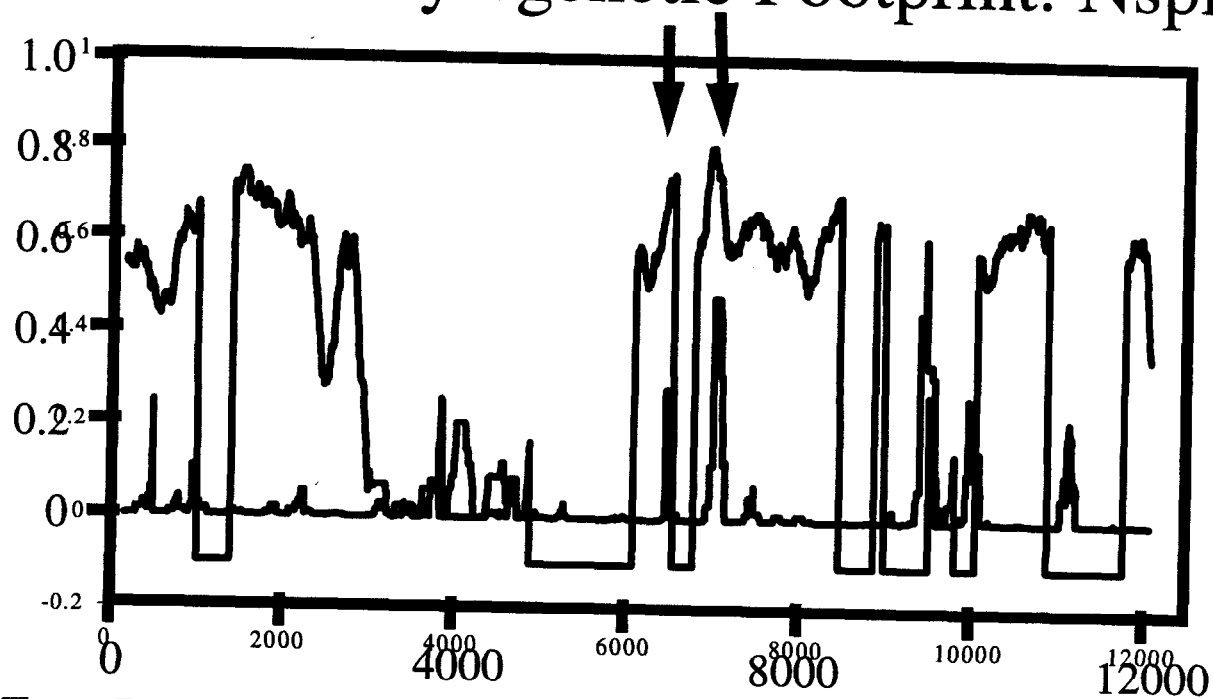


Corroborative Evidence

- ◆ Beneficial to identify regions as interesting using alternative approach
- ◆ Phylogenetic Footprinting
 - Conserved sequence features suggest function
- ◆ Limited to genes with sequence available from multiple species




LRA and Phylogenetic Footprint: Nspl-1



Bottom line

- If a muscle gene, $p \sim 0.85$ to find it
- If a hit, $p \sim 0.25$ of muscle gene;
Better if conserved between species
- The big challenge is to generalize to tissues/states with much less data

Concluding Thoughts

- ◆ Regulatory Modules Important for Computational Specificity
 - ◆ Logistic Regression Analysis is an Effective Tool
 - ◆ Phylogenetic Footprinting Complements Analysis
 - ◆ Extension of Approach Challenging
- 

Acknowledgments

- Promoter recognition review in collaboration with A. Hatzigeorgiou
- Muscle transcription work in collaboration with W. Wasserman
- Funded by NHGRI, SB, and Synaptic LTD